Bayes Optimal Kernel Discriminant Analysis

Di You and Aleix M. Martinez
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH 43210, USA
youd@ece.osu.edu aleix@ece.osu.edu

Abstract

Kernel methods provide an efficient mechanism to derive nonlinear algorithms. In classification problems as well as in feature extraction, kernel-based approaches map the originally nonlinearly separable data into a space of intrinsically much higher dimensionality where the data is linearly separable and can be readily classified with existing and efficient linear methods. For a given kernel function, the main challenge is to determine the parameters of the kernel which map the original nonlinear problem to a linear one. This paper derives a Bayes optimal criterion for the selection of the kernel parameters in discriminant analysis. Our criterion selects the kernel parameters that maximize the (Bayes) classification accuracy in the kernel space. We also show how we can use the same criterion to do subclass selection in the kernel space for problems with multimodal class distributions. Extensive experimental evaluation demonstrates the superiority of the proposed criterion over the state of the art.

1. Introduction

Kernel mapping has become one of the most popular approaches to perform nonlinear feature extraction and classification with many applications, such as, object and face recognition, handwritten text analysis, and classification of specimens in paleontology to name but a few [17, 22, 21, 8]. Using a kernel function, the original data is implicitly mapped to a very high or even infinite dimensional space where the data is linearly separable. Then, any efficient linear method can be employed in this so called kernel space. Since the mapping is intrinsic, one does not need to work with an explicit mapping function. Instead, one can employ the kernel trick [18, 17], resulting in a space of the same dimensionality as that of the input representation while still eliminating the nonlinearity of the data.

Kernel Discriminant Analysis (KDA) [1, 15] is one of the most common techniques used in feature extraction and classification. KDA is a kernel extension of Linear Discriminant Analysis (LDA) [5]. KDA aims to maximize the between-class scatter and minimize the within-class scatter of the data simultaneously. While LDA attempts to do so in the original space, KDA does this in the kernel space. If the class distributions in the original space are homoscedastic (*i.e*., identical covariance matrices), then LDA will yield the Bayes optimal solution, that is, the hyperplane separating the class distributions in the LDA subspace will have the smallest possible error (also known as the Bayes error). Unfortunately, the class distributions are rarely homoscedastic and the optimal solution cannot be attained. To resolve this problem, we can employ the kernel trick.

For a given kernel function, the goal is to determine the kernel parameters to achieve Bayes optimal classification in the kernel space. Care needs to be taken in this selection process. If the kernel parameter makes the model too complex, an over-fitting problem to the training data may result. If the model is made too simple, it may underfit the data and will thus not effectively capture the underlying structure of the data.

The classical approach to determine the kernel parameters is cross-validation (CV). In this technique, the parameters which minimize the validation error are selected via an exhaustive search. However, this method is computationally expensive and merely selects the parameters from a set of prespecified discrete values. To avoid these two drawbacks, [19, 20] define a criterion, which maximizes the between-class scatter and minimizes the within-class scatter in the kernel space, to optimize the kernel parameters. Since the idea is similar to between-within class ratio of LDA [5], we will refer to this as the Fisher criterion. This criterion maximizes the class separability in the kernel space, and it generally obtains higher classification accuracies than CV. A similar idea is developed in [9], where the Fisher criterion is reformulated as a convex optimization problem and then used to find a solution over a convex set of kernels. Alternatively, [3] defines the concept of kernel alignment to capture the agreement between a kernel and the target data. It is shown how this measure can be used to optimize the kernel. Interestingly, [20] shows that this kernel-target alignment criterion is equivalent to maximizing the between-class scatter, provided that the kernel matrix has been centralized and normalized by its Frobenius norm. These approaches will thus be grouped within the idea of the Fisher criterion. The major drawback with these criteria is that they are only based on the measures of class separability. Note that the measure for the class separability is not always related to the classification error. For example, since the Fisher criterion is based on a least-squares formulation [8], this can easily over-weight the influence of the classes that are farthest apart (*i.e.*, well separated) [12].

In the present paper, we derive a Bayes optimal criterion for selecting the parameter in discriminant analysis. To achieve this, we define a function measuring the Bayes accuracy (*i.e.*, one minus the Bayes error) in the kernel space. We then show how this function can be efficiently maximized using gradient ascent. It should be emphasized that this objective function directly minimizes the classification error, which makes the proposed criterion very powerful. We will also illustrate how we can employ the same criterion for the selection of other parameters in discriminant analysis. In particular, we demonstrate the uses of the derived criterion in the selection of the kernel parameters and the number of subclasses in Kernel Subclass Discriminant Analysis (KSDA), a kernel version of Subclass Discriminant Analysis (SDA) [23]. Extensive experiments demonstrate that our criterion generally yields higher classification accuracies than others.

2. Background Formulation and Notation

LDA simultaneously maximizes the between-class scatter and minimizes the total scatter (*i.e.*, covariance matrix) [5, 14], given by $\mathbf{S}_B = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j (\mu_i - \mu_j)(\mu_i - \mu_j)^T$ and $\mathbf{\Sigma}_X = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$, where $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ are the n training samples, $\mathbf{x}_i \in \mathbb{R}^p$, $\mu = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$ is the sample mean, $\mu_i = n_i^{-1}\sum_{j=1}^{n_i}\mathbf{x}_{ij}$ is the sample mean of class i, $\mathbf{x}_{ij}$ is the j^{th} sample of class i, n_i is the number of samples in class i, C is the number of classes, and $p_i = n_i/n$ is the prior of class i. LDA's solution is then given by the generalized eigenvalue decomposition equation $\mathbf{\Sigma}_X^{-1}\mathbf{S}_B\mathbf{V} = \mathbf{V}\Lambda$, where the columns of $\mathbf{V}$ are the eigenvectors, and Λ is a diagonal matrix of corresponding eigenvalues. If the class distributions are homoscedastic, the subspace given by the eigenvectors with nonzero eigenvalue of this equation yield the Bayes optimal solution.

In general, the class distributions are not homoscedastic. In such cases, the solution given by LDA is biased toward those classes that are furthest apart. To see this, note that LDA is based on least-squares (*i.e.*, an eigenvalue decomposition defined to solve a system of homogeneous equations [8]). Thus, the LDA solution tends to over-weight the classes that were already well-separated in the original space. In order to downplay the roles of the class distributions that are farthest apart, [12] introduces a weighted version of $\mathbf{S}_B$, defined as

$$\mathbf{\Sigma}_B = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij})(\mu_i - \mu_j)(\mu_i - \mu_j)^T, \quad (1)$$

where $\Delta_{ij}^2 = (\mu_i - \mu_j)^T\mathbf{\Sigma}_X^{-1}(\mu_i - \mu_j)$ is the Mahalanobis distance between classes i and j, $\omega : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ is a weighting function, $\omega(\Delta_{ij}) = \frac{1}{2\Delta_{ij}^2} erf(\frac{\Delta_{ij}}{2\sqrt{2}})$, and $erf(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the error function.

One advantage of (1) is that it is related to the mean pairwise Bayes accuracy [12] (*i.e.*, one minus the Bayes error), since

$$J(\mathbf{L}) = \sum_{m=1}^{d}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij}) tr(\mathbf{e}_m^T\mathbf{\Sigma}_{ij}\mathbf{e}_m), \quad (2)$$

where $\mathbf{L} = (\mathbf{e}_1, ..., \mathbf{e}_d)$ is the eigenvector matrix of $\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij})\mathbf{\Sigma}_{ij}$, $\mathbf{\Sigma}_{ij} = (\mu_i - \mu_j)(\mu_i - \mu_j)^T$ are the pairwise class distances, and, for simplicity, we have assumed $\mathbf{\Sigma}_X = \mathbf{I}_p$, $\mathbf{I}_p$ an identity matrix with dimension $p \times p$.

3. Bayes Optimal Criterion

3.1. Bayes accuracy in the kernel space

As mentioned above, (2) is proportional to the Bayes accuracy and as such it can be employed to improve LDA [12]. We want to derive a similar function for its use in the kernel space.

Let $\phi(.) : \mathbb{R}^p \rightarrow \mathcal{F}$ be a function defining the kernel map. We also assume the data has already been whitened in the kernel space. Denote the data matrix in the kernel space $\Phi(\mathbf{X})$, where $\Phi(\mathbf{X}) = (\phi(\mathbf{x}_{11}), \ldots, \phi(\mathbf{x}_{in_i}), \ldots, \phi(\mathbf{x}_{Cn_C}))$. The kernel matrix is given by $\mathbf{K} = \Phi(\mathbf{X})^T\Phi(\mathbf{X})$.

Using this notation, the covariance matrix in the kernel space can be written as $\mathbf{\Sigma}_X^{\Phi} = n^{-1}\Phi(\mathbf{X})(\mathbf{I}_n - \mathbf{P}_n)\Phi(\mathbf{X})^T$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix, and $\mathbf{P}_n$ is a $n \times n$ matrix with all elements equal to $1/n$. The whitened data matrix $\tilde{\Phi}(\mathbf{X})$ is now given by $\tilde{\Phi}(\mathbf{X}) = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^{\Phi^T}\Phi(\mathbf{X})$, where $\mathbf{\Lambda}$ and $\mathbf{V}^{\Phi}$ are the eigenvalue and eigenvector matrices given by $\mathbf{\Sigma}_X^{\Phi}\mathbf{V}^{\Phi} = \mathbf{V}^{\Phi}\mathbf{\Lambda}$. We know from the Representer's Theorem [18] that a projection vector lies in the span of the samples in the kernel space $\Phi(\mathbf{X})$, *i.e.*, $\mathbf{V}^{\Phi} = \Phi(\mathbf{X})\mathbf{\Gamma}$, where $\mathbf{\Gamma}$ is a corresponding coefficient matrix. Thus, we have

$$\begin{aligned}\tilde{\Phi}(\mathbf{X}) &= \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{V}^{\Phi^T}\Phi(\mathbf{X}) \\ &= \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Gamma}^T\Phi(\mathbf{X})^T\Phi(\mathbf{X}) = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{\Gamma}^T\mathbf{K},\end{aligned}$$

where $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ can be calculated from a generalized eigenvalue decomposition problem $\mathbf{N\Gamma} = \mathbf{K\Gamma\Lambda}$, with $\mathbf{N} = n^{-1}\mathbf{K}(\mathbf{I}_n - \mathbf{P}_n)\mathbf{K}$. With this trick, we transform the kernel covariance matrix $\mathbf{\Sigma}_X^{\Phi}$ into the identity matrix.

Next, define the mean of class i in the kernel space as

$$\mu_i^{\phi} = \tilde{\Phi}(\mathbf{X}_i)\mathbf{1}_i, \tag{3}$$

where $\tilde{\Phi}(\mathbf{X}_i) = (\tilde{\phi}(\mathbf{x}_{i1}), \ldots, \tilde{\phi}(\mathbf{x}_{in_i}))$, and $\mathbf{1}_i$ is a $n_i \times 1$ vector with all elements equal to $1/n_i$. Let $\tilde{\mathbf{K}}_i = \tilde{\Phi}(\mathbf{X})^T\tilde{\Phi}(\mathbf{X}_i)$ denote the subset of the whitened kernel matrix for the samples in class i.

Combining the above results, we can define the Bayes accuracy in the kernel space as

$$Q(\phi) = \sum_{m=1}^{d}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij}^{\Phi}) \mathbf{e}_m^{\phi^T} \mathbf{S}_{ij}^{\Phi} \mathbf{e}_m^{\phi}, \tag{4}$$

where $\mathbf{e}_1^{\phi}, ..., \mathbf{e}_d^{\phi}$ are the eigenvectors of the weighted kernel between-class scatter matrix

$$\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij}^{\Phi}) \mathbf{S}_{ij}^{\Phi},$$

$\mathbf{S}_{ij}^{\Phi} = (\mu_i^{\phi} - \mu_j^{\phi})(\mu_i^{\phi} - \mu_j^{\phi})^T$, the Mahalanobis distance Δ_{ij}^{Φ} in the whitened kernel space becomes the Euclidean distance,

$$\begin{aligned} \Delta_{ij}^{\Phi^2} &= (\mu_i^{\phi} - \mu_j^{\phi})^T(\mu_i^{\phi} - \mu_j^{\phi}) \\ &= (\tilde{\Phi}(\mathbf{X}_i)\mathbf{1}_i - \tilde{\Phi}(\mathbf{X}_j)\mathbf{1}_j)^T(\tilde{\Phi}(\mathbf{X}_i)\mathbf{1}_i - \tilde{\Phi}(\mathbf{X}_j)\mathbf{1}_j) \\ &= \mathbf{1}_i^T\tilde{\mathbf{K}}_{ii}\mathbf{1}_i - 2\mathbf{1}_i^T\tilde{\mathbf{K}}_{ij}\mathbf{1}_j + \mathbf{1}_j^T\tilde{\mathbf{K}}_{jj}\mathbf{1}_j, \end{aligned} \tag{5}$$

and $\tilde{\mathbf{K}}_{ij} = \tilde{\Phi}(\mathbf{X}_i)^T\tilde{\Phi}(\mathbf{X}_j)$ is the subset of the kernel matrix for the samples in class i and j.

From the Representer's Theorem [18], we know that $\mathbf{e}_i^{\phi} = \tilde{\Phi}(\mathbf{X})\mathbf{u}_i$, where $\mathbf{u}_i$ is a coefficient vector. Then, using (3) we have $\mathbf{e}_m^{\phi^T}\mathbf{S}_{ij}^{\Phi}\mathbf{e}_m^{\phi} = \mathbf{u}_m^T\mathcal{S}_{ij}\mathbf{u}_m$, where $\mathcal{S}_{ij} = (\tilde{\mathbf{K}}_i\mathbf{1}_i - \tilde{\mathbf{K}}_j\mathbf{1}_j)(\tilde{\mathbf{K}}_i\mathbf{1}_i - \tilde{\mathbf{K}}_j\mathbf{1}_j)^T$, and $\mathbf{u}_1, \ldots, \mathbf{u}_d$ are the eigenvectors of $\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij}^{\Phi})\mathcal{S}_{ij}$. Therefore, criterion (4) can be rewritten as

$$Q(\phi) = \sum_{m=1}^{d}\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij}^{\Phi}) \mathbf{u}_m^T \mathcal{S}_{ij} \mathbf{u}_m. \tag{6}$$

By maximizing $Q(\phi)$, we favor a kernel representation where the sum of pairwise Bayes accuracies is maximized. The optimal kernel function, ϕ^*, is given by

$$\phi^* = \arg\max_{\phi} Q(\phi).$$

We will refer to the derived criterion given in (6) as Kernel Bayes Accuracy (KBA) criterion.

3.2. Kernel parameters with gradient ascent

The first application of the above derived criterion is in determining the value of the parameters of a kernel function. For example, if we are given the Radial Basis Function (RBF) kernel, $k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$, our goal is to determine an appropriate value of the variance σ.

To determine our solution, we employ a quasi-Newton method with a Broyden-Fletcher-Goldfarb-Shanno Hessian update [4]. The main advantage of this method is that it has a fast converge and does not require the calculation of the Hessian matrix. Instead, the Hessian is updated by analyzing the gradient vectors.

To compute the derivative of our criterion, note that (6) can be rewritten as

$$\begin{aligned} Q(\phi) &= tr(\sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij}^{\Phi})\mathcal{S}_{ij}) \\ &= \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \omega(\Delta_{ij}^{\Phi}) tr(\mathcal{S}_{ij}). \end{aligned}$$

Taking the partial derivative with respect to σ in the RBF kernel, we have $\frac{\partial Q(\phi)}{\partial\sigma} = \sum_{i=1}^{C-1}\sum_{j=i+1}^{C} p_i p_j \left[\frac{\partial\omega(\Delta_{ij}^{\Phi})}{\partial\sigma} tr(\mathcal{S}_{ij}) + \omega(\Delta_{ij}^{\Phi})\frac{\partial tr(\mathcal{S}_{ij})}{\partial\sigma}\right]$.

Denote the partial derivative of an $m \times n$ matrix $\mathcal{K}$ with respect to σ as $\frac{\partial\mathcal{K}}{\partial\sigma} = \left[\frac{\partial\mathcal{K}_{ij}}{\partial\sigma}\right]_{i=1,\ldots,m,j=1,\ldots,n}$, with $\frac{\partial\mathcal{K}_{ij}}{\partial\sigma} = \frac{\partial k(x_i,x_j)}{\partial\sigma} = \frac{\|x_i - x_j\|^2}{\sigma^3}\exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$. Then $\frac{\partial\omega(\Delta_{ij}^{\Phi})}{\partial\sigma} = -\frac{erf(\Delta_{ij}^{\Phi}/2\sqrt{2})}{\Delta_{ij}^{\Phi^3}}\frac{\partial\Delta_{ij}^{\Phi}}{\partial\sigma} + \frac{\exp(-\Delta_{ij}^{\Phi^2}/8)}{2\sqrt{2\pi}\Delta_{ij}^{\Phi^2}}\frac{\partial\Delta_{ij}^{\Phi}}{\partial\sigma}$, where $\frac{\partial\Delta_{ij}^{\Phi}}{\partial\sigma} = \frac{1}{2\Delta_{ij}^{\Phi}}(\mathbf{1}_i^T\frac{\partial\tilde{\mathbf{K}}_{ii}}{\partial\sigma}\mathbf{1}_i - 2\mathbf{1}_i^T\frac{\partial\tilde{\mathbf{K}}_{ij}}{\partial\sigma}\mathbf{1}_j + \mathbf{1}_j^T\frac{\partial\tilde{\mathbf{K}}_{jj}}{\partial\sigma}\mathbf{1}_j)$.

Finally, $\frac{\partial tr(\mathcal{S}_{ij})}{\partial\sigma} = \frac{\partial(\tilde{\mathbf{K}}_i\mathbf{1}_i - \tilde{\mathbf{K}}_j\mathbf{1}_j)^T(\tilde{\mathbf{K}}_i\mathbf{1}_i - \tilde{\mathbf{K}}_j\mathbf{1}_j)}{\partial\sigma} = \mathbf{1}_i^T\frac{\partial\tilde{\mathbf{K}}_i^T}{\partial\sigma}\tilde{\mathbf{K}}_i\mathbf{1}_i + \mathbf{1}_i^T\tilde{\mathbf{K}}_i^T\frac{\partial\tilde{\mathbf{K}}_i}{\partial\sigma}\mathbf{1}_i - 2\mathbf{1}_j^T\frac{\partial\tilde{\mathbf{K}}_j^T}{\partial\sigma}\tilde{\mathbf{K}}_i\mathbf{1}_i - 2\mathbf{1}_j^T\tilde{\mathbf{K}}_j^T\frac{\partial\tilde{\mathbf{K}}_i}{\partial\sigma}\mathbf{1}_i + \mathbf{1}_j^T\frac{\partial\tilde{\mathbf{K}}_j^T}{\partial\sigma}\tilde{\mathbf{K}}_j\mathbf{1}_j + \mathbf{1}_j^T\tilde{\mathbf{K}}_j^T\frac{\partial\tilde{\mathbf{K}}_j}{\partial\sigma}\mathbf{1}_j$.

3.3. Subclass extension

Another application of the derived KBA criterion is in determining the number of subclasses in Subclass Discriminant Analysis (SDA) [23] and its kernel extension. KDA assumes that each class has a single Gaussian distribution in the kernel space. However, this may be too restrictive since it is usually difficult to find a kernel representation where the class distributions are single Gaussians. In order to relax this assumption, we can describe each class using a mixture of Gaussians. Using this idea, we can reformulate (6) as

$$\begin{aligned} Q^{sub}(\phi, H_1, \ldots, H_C) = &\sum_{m=1}^{d}\sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{k=i+1}^{C}\sum_{l=1}^{H_k} p_{ij}p_{kl} \\ &\omega(\Delta_{ij,kl}^{\Phi})\mathbf{u}_m^T\mathcal{S}_{ij,kl}\mathbf{u}_m, \end{aligned} \tag{7}$$

where H_i is the number of subclasses in class i, $\mathbf{u}_1, ..., \mathbf{u}_d$ are d eigenvectors of the kernel version of the weighted between-*sub*class scatter matrix

$$\sum_{i=1}^{C-1}\sum_{j=1}^{H_i}\sum_{k=i+1}^{C}\sum_{l=1}^{H_k} p_{ij}p_{kl}\omega(\Delta_{ij,kl}^{\Phi})\mathcal{S}_{ij,kl},$$

$\mathcal{S}_{ij,kl} = (\mathbf{M}_{ij}\mathbf{1}_{ij} - \mathbf{M}_{kl}\mathbf{1}_{kl})(\mathbf{M}_{ij}\mathbf{1}_{ij} - \mathbf{M}_{kl}\mathbf{1}_{kl})^T$, $\mathbf{M}_{ij} = \tilde{\Phi}(\mathbf{X})^T\tilde{\Phi}(\mathbf{X}_{ij})$, $\tilde{\Phi}(\mathbf{X}_{ij}) = (\tilde{\phi}(\mathbf{x}_{ij1}), \ldots, \tilde{\phi}(\mathbf{x}_{ijn_{ij}}))$, $\mathbf{x}_{ijk}$ is the k^{th} sample of subclass j in class i, $\mathbf{1}_{ij}$ is a $n_{ij} \times 1$ vector with all elements equal to $1/n_{ij}$, and n_{ij} the number of samples in the j^{th} subclass of class i. Note that in the above equation, the whitened Mahalanobis distance is given by

$$\begin{aligned}\Delta_{ij,kl}^{\Phi^2} &= (\mu_{ij}^{\phi} - \mu_{kl}^{\phi})^T(\mu_{ij}^{\phi} - \mu_{kl}^{\phi}) \\ &= \mathbf{1}_{ij}^T\tilde{\mathbf{K}}_{ij,ij}\mathbf{1}_{ij} - 2\mathbf{1}_{ij}^T\tilde{\mathbf{K}}_{ij,kl}\mathbf{1}_{kl} + \mathbf{1}_{kl}^T\tilde{\mathbf{K}}_{kl,kl}\mathbf{1}_{kl},\end{aligned}$$

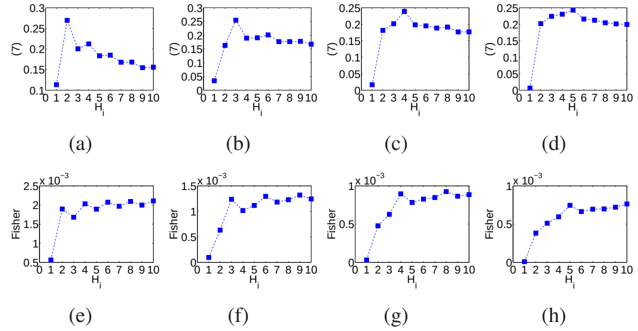
where $\tilde{\mathbf{K}}_{ij,kl} = \tilde{\Phi}(\mathbf{X}_{ij})^T\tilde{\Phi}(\mathbf{X}_{kl})$. The optimal kernel function and subclass divisions are given by

$$\phi^*, H_1^*, \ldots, H_C^* = \arg\max_{\phi, H_1, \ldots, H_C} Q^{sub}(\phi, H_1, \ldots, H_C).$$

3.4. Optimal subclass discovery

In KSDA we are simultaneously optimizing the kernel parameter and the number of subclasses. It is in fact advantageous to do so, because it will allow us to find the Bayes optimal solution when the classes need to be described with a mixture of Gaussians in the kernel space. Furthermore, we can automatically determine the underlying structure of the data. This last point is important in many applications. We illustrate this with a set of examples.

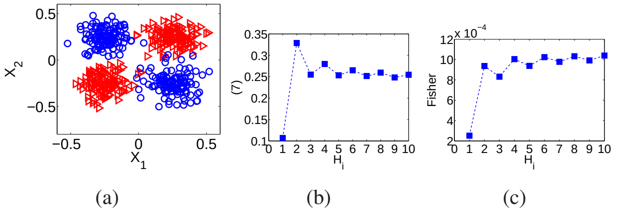
In our case study, we generated a set of 120 samples for each of the two classes. Each class was represented by a mixture of two Gaussians, with mean and diagonal covariance randomly initialized. Then, (7) was employed to determine the appropriate number of subclasses and parameter of the RBF kernel. This process was repeated 100 times, each with a different random initialization of the means and covariances. The average of the maxima of (7) for each value of H_i (with $H_1 = H_2$) are shown in Fig. 1(a). We see that the derived criterion is on average higher for the correct number of subclasses. We then repeated the process described in this paragraph for the cases of 3, 4 and 5 subclasses per class. The results are in Fig. 1(b-d). Again, the maximum of (7) corresponds to the correct number of subclasses. Therefore, the proposed criterion can generally be efficiently employed to discover the underlying structure of the data. For comparison, in Fig. 1(e-h) we show the plots of the Fisher criterion described earlier. We see that this criterion does not recover the correct number of subclasses and is generally monotonically increasing, thus, tending to select large values for H_i. This is because the Fisher criterion maximizes the between-subclass scatter and, generally, the larger H_i, the larger the scatter.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1. Comparative results between the (a-d) KBA and (e-h) Fisher criteria. The true underlying number of subclasses per class are (a,e) 2, (b,f) 3, (c,g) 4, and (d,h) 5. The x-axis specifies the number of subclasses H_i. The y-axis shows the value of the criterion given in (7) in (a-d) and of the Fisher criterion in (e-h).



(a) (b) (c)

Figure 2. (a) The classical XOR classification problem. (b) Plot of the KBA criterion versus H_i. (c) Plot of the Fisher criterion.

As a more challenging case, we also consider the well-known XOR data classification problem, Fig. 2(a). The values of (7) for different H_i are plotted in Fig. 2(b) and those of the Fisher criterion in (c). Once more, we see that the KBA criterion is capable of accurately recovering the number of subclasses, whereas the Fisher criterion is not.

4. Experimental results

We now present results on a variety of classification problems. We use the derived criterion in KDA and KSDA. We use the notation KDA$_K$ and KSDA$_K$ to indicate that the KBA criterion was used to optimize the parameters. We provide comparative results with CV and the Fisher criterion, denoted with a CV and F subscript, respectively. We also provide comparative results against Principal Component Analysis (PCA), LDA, accurate Pairwise Accuracy Criteria (aPAC) [12], Nonparametric DA (NDA) [7], SDA [23], Regularized DA (RDA) [6], Kernel PCA (KPCA) and Heteroscedastic LDA (HLDA) [11].

4.1. Databases

The first five data-sets we will use are from the UCI repository [2]. The Monks problem goal is to learn to discriminate between two different postures of a robot. Three different case scenarios are considered, denoted Monk 1, 2, and 3. The Ionosphere set corresponds to satellite imaging for the detection of two classes (structure or not) in the ground. And, the NIH Pima set is used to detect diabetes from a set of eight measurements.

The next database we will use is the ETH-80 [10], which includes a total of $3,280$ images of 8 categories. Each category includes 10 objects and each of these 80 objects has been photographed from 41 different positions. All the images are resized to 25×30 pixels. The pixel values in their vector form (in $\mathbb{R}^{750}$) are used to define the original feature space. As it is typical in this database, we will use the leave-one-object-out test. In this test, we use the 41 images of one of the 80 objects for testing and the images of the remaining 79 objects are employed for training. Since there are 80 ways of selecting the testing group, we test them all and calculate the average recognition rate.

We will also use 100 randomly selected subjects from the AR face database [13]. All images were aligned with respect to the eyes, mouth and jaw line before cropping and resizing them to a standard size of 29×21 pixels. This database contains images of two different sessions, taken two weeks apart. We will use the images in the first session for training and those in the second for testing.

The final data-set we will use is the Sitting Posture Distribution Maps (SPDM) [24]. In this data set, samples were collected using a chair equipped with a pressure sensor sheet located on the sit-pan and back-rest. The pressure maps provide a total of $1,280$ pressure values from 50 individuals. Each participant provides five samples of each of the ten different postures. Our goal is to classify each of the samples into one of the ten sitting postures. We randomly selected 3 samples from each individual and posture for training, and used the rest for testing.

4.2. Results and Analysis

The linear and nonlinear feature extraction methods described earlier are used to find an appropriate low-dimensional representation of the data. Here, we use the classical RBF kernel defined earlier. In this low-dimensional space, one can use a variety of classification techniques. In this section, we provide successful classification results using three methods: the classical nearest neighbor (NN) classifier, the extension of K-NN defined in [16], and a linear Support Vector Machines (SVM). The recognition results are shown in Tables 1-3.

DATA SET	KSDA$_K$	KSDA$_F$	KSDA$_{CV}$	KDA$_K$	KDA$_F$	KDA$_{CV}$	KPCA
ETH-80	**84.6***	73.6	76.8	**84.6***	81.0	71.6	62.2
AR DATABASE	**88.2***	78.3	84.2	86.1	**87.5**	84.2	42.5
SPDM	**84.3***	80.1	83.7	**84.3***	**84.2**	83.3	75.0
MONK1	**88.0**	84.5	87.5	87.3	**89.6***	83.1	90.3*
MONK2	**82.9***	**83.1***	75.7	**82.9***	75.2	70.1	68.3
MONK3	**94.2***	87.7	89.8	**92.6**	88.0	82.4	87.8
IONOSPHERE	93.0	84.8	**94.0***	**89.1**	86.5	80.8	89.4
PIMA	73.2	73.8	**76.8***	**76.2***	69.8	72.6	56.0

DATA SET	PCA	LDA	NDA	APAC	HLDA	RDA	SDA
ETH-80	64.3	64.3	59.8	73.6	56.5	71.6	70.6
AR DATABASE	58.6	77.7	77.0	59.1	67.5	78.6	77.7
SPDM	81.5	66.5	48.8	81.1	65.3	59.5	66.1
MONK1	81.3	69.0	68.3	81.0	84.2	72.0	75.7
MONK2	66.7	67.4	82.6	79.6	83.6	60.0	67.4
MONK3	87.3	70.6	83.6	88.4	84.5	86.3	85.9
IONOSPHERE	92.1	74.8	88.8	92.1	88.7	82.8	93.4*
PIMA	64.3	57.7	69.1	62.5	68.5	66.7	57.7

Table 1. Recognition rates (%) with nearest neighbor. Bold numbers specify the top recognition obtained with the three criteria in KSDA and KDA. An asterisk specifies a statistical significance on the highest recognition rate.

DATA SET	KSDA$_K$	KSDA$_F$	KSDA$_{CV}$	KDA$_K$	KDA$_F$	KDA$_{CV}$	KPCA
ETH-80	**84.6***	73.9	76.4	**84.6***	82.8	72.9	60.3
AR DATABASE	**89.6***	78.5	85.1	**87.5**	86.7	85.1	49.5
SPDM	**84.9***	75.3	83.9	**84.9***	83.4	82.6	75.0
MONK1	**88.0***	76.6	82.9	87.3	87.7	**88.7***	77.3
MONK2	**82.9***	77.5	75.7	**82.9***	**82.9***	78.5	58.6
MONK3	90.5	83.3	86.3	**92.6**	**92.4**	91.2	91.2
IONOSPHERE	**92.8***	84.8	86.1	**89.1**	86.8	86.8	82.1
PIMA	**78.6***	76.8	76.2	**76.2**	73.0	69.0	60.7

DATA SET	PCA	LDA	NDA	APAC	HLDA	RDA	SDA
ETH-80	67.1	64.3	63.5	71.2	59.1	71.6	72.3
AR DATABASE	44.5	70.9	77.3	60.2	67.5	78.6	70.9
SPDM	77.0	56.2	50.2	81.2	53.4	59.5	69.5
MONK1	78.2	67.4	77.8	69.4	71.5	72.0	79.2
MONK2	56.7	70.6	70.6	70.4	58.3	60.0	70.6
MONK3	89.7	70.8	91.9	89.6	93.8*	86.3	90.5
IONOSPHERE	82.1	74.8	83.4	91.1	94.0*	82.8	89.4
PIMA	70.2	57.7	70.2	63.8	72.6	66.7	57.7

Table 2. Recognition rates (%) with the classification method of [16].

DATA SET	KSDA$_K$	KSDA$_F$	KSDA$_{CV}$	KDA$_K$	KDA$_F$	KDA$_{CV}$	KPCA
ETH-80	**84.2***	73.6	77.4	**84.2***	82.2	71.3	65.3
AR DATABASE	**86.7***	79.6	83.1	85.3	**86.7***	83.1	42.1
SPDM	**84.3***	**84.6***	82.3	**84.3***	83.6	82.6	66.7
MONK1	87.3	**88.2**	86.1	87.3	**89.7***	86.1	88.4*
MONK2	**82.9***	81.5	73.8	**82.9***	75.2	75.1	50.0
MONK3	93.5	91.9	**94.4***	**91.9**	89.1	81.5	94.4
IONOSPHERE	92.6	86.1	**96.7***	**89.1**	86.1	82.1	82.1
PIMA	**79.8***	78.6	**79.8***	**77.4**	75.0	72.8	64.3

DATA SET	PCA	LDA	NDA	APAC	HLDA	RDA	SDA
ETH-80	60.1	65.3	61.8	68.4	68.4	71.6	67.8
AR DATABASE	66.7	79.3	69.7	67.2	70.1	78.6	79.3
SPDM	76.5	50.3	49.0	82.1	69.3	59.5	69.0
MONK1	67.8	65.6	66.4	67.8	68.5	72.0	66.7
MONK2	67.1	67.1	67.5	65.6	67.1	60.0	67.1
MONK3	81.3	63.9	83.3	80.6	81.9	86.3	84.7
IONOSPHERE	84.8	84.8	88.1	93.4	93.4	82.8	90.1
PIMA	68.6	64.9	76.8	77.4	76.2	66.7	64.9

Table 3. Recognition rates (%) with linear SVM.

From these results, it is clear that, on average, the derived KBA criterion achieves higher classification rates than the Fisher criterion and CV. As expected, KSDA generally yields superior results than KDA. This is due to the added flexibility on modeling the underlying class distributions in the kernel space provided by KSDA. To illustrate the effectiveness of the proposed criterion in KSDA, we show the smoothness of the function optimized by the criterion in Fig. 3 for four of the data-sets. Note how these functions can be readily optimized using gradient ascent. It is also interesting to note that the optimal value of σ remains relatively constant for different values of H_i. This smoothness in the change of the criterion is what allows to find the global optimum efficiently.

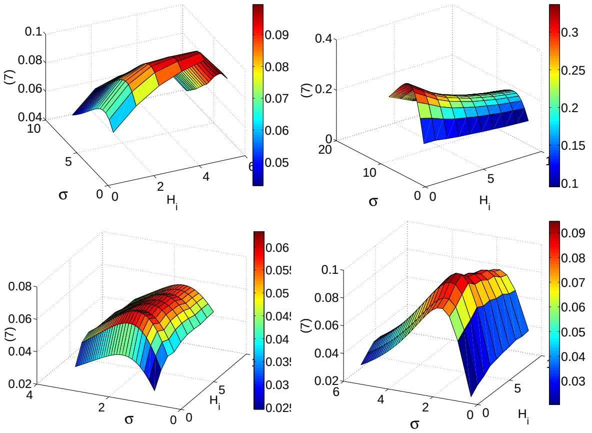


Figure 3. Plots of the value of the derived criterion as a function of the kernel parameter and the number of subclasses. From left to right and top to bottom: AR, ETH-80, Monk 1, and Ionosphere databases.

5. Conclusions

We have derived a Bayes optimal criterion for the selection of the kernel parameters in KDA and the number of subclasses and kernel parameters in KSDA. The derived function computes the Bayes accuracy, defined as one minus the Bayes error, in the kernel space. Thus, the goal is to find that kernel representation where the highest classification accuracy is achieved. We have also shown how this criterion can be efficiently optimized using gradient ascent without the need to explicitly compute the Hessian. Extensive experimental results on a number of databases shows that the derived approach yields superior classification results to those given by existing algorithms. Moreover, we have demonstrated that, when used in KSDA, the proposed criterion can accurately recover the underlying structure of the class distributions.

Acknowledgments

This research was partially funded by the National Science Foundation under grant IIS 0713055.

References

[1] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2835–2404, 2000. 1

[2] C. L. Blake and C. J. Merz. UCI repository of machine learning databases. University of California, Irvine, http ://www.ics.uci.edu/mlearn/MLRepository.html, 1998. 5

[3] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor. On kernel target alignment. In *Proc. Advances in Neural Information Processing Systems*, pages 367–373, 2001. 1

[4] J. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall, 1983. 3

[5] R. A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386, 1938. 1, 2

[6] J. H. Friedman. Regularized discriminant analysis. *J. Am. Stat. Assoc.*, 84:165–175, 1989. 4

[7] K. Fukunaga and J. Mantock. Nonparametric discriminant analysis. *IEEE Trans. PAMI*, 5:671–678, 1983. 4

[8] O. C. Hamsici and A. M. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Trans. PAMI*, 30:647–657, 2008. 1, 2

[9] S. Kim, A. Magnani, and S. Boyd. Optimal kernel selection in kernel fisher discriminant analysis. In *Int. Conf. Machine Learning*, pages 465–472, 2006. 1

[10] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *Proc. IEEE Conf. CVPR*, 2003. 5

[11] M. Loog and R. P. W. Duin. Linear dimensionality reduction via a heteroscedastic extension of lda: The chernoff criterion. *IEEE Trans. PAMI*, 26(6):732–739, 2004. 4

[12] M. Loog, R. P. W. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Trans. PAMI*, 23(7):762–766, 2001. 2, 4

[13] A. M. Martinez and R. Benavente. *The AR Face Database*. CVC Technical Report No. 24, June, 1998. 5

[14] A. M. Martinez and M. Zhu. Where are linear feature extraction methods applicable? *IEEE Trans. PAMI*, 27(12):1934–1944, 2005. 2

[15] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. Fisher discriminant analysis with kernels. In *Proceedings of IEEE Neur. Net. Sig. Proc. Workshop*, pages 41–48, 1999. 1

[16] O. Pujol and D. Masip. Geometry-based ensembles: Towards a structural characterization of the classification boundary. *IEEE Trans. PAMI*, 31(6):1140–1146, 2009. 5

[17] B. Schlkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2001. 1

[18] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. 1, 2, 3

[19] L. Wang, K. Chan, P. Xue, and L. Zhou. A kernel-induced space selection approach to model selection in klda. *IEEE Trans. Neural Networks*, 19:2116–2131, 2008. 1

[20] H. Xiong, M. Swamy, and M. Ahmad. Optimizing the kernel in the empirical feature space. *IEEE Trans. on Neural Networks*, 16(2):460–474, 2005. 1

[21] J. Yang, A. F. Frangi, J. Yang, D. Zhang, and Z. Jin. KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition. *IEEE Trans. PAMI*, 27(2):230–244, 2005. 1

[22] M.-H. Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *FGR '02: Proc. IEEE Int. Conf. on Automatic Face and Gesture Rec*, 2002. 1

[23] M. Zhu and A. M. Martinez. Subclass discriminant analysis. *IEEE Trans. PAMI*, 28(8):1274–1286, 2006. 2, 3, 4

[24] M. Zhu and A. M. Martinez. Pruning noisy bases in discriminant analysis. *IEEE Trans. Neural Networks*, 19(1):148–157, 2008. 5